E-006

表形式における情報の分類手法
The Method Of Classifying Information In Tabular Form

目片 亮太郎 †　　土屋 誠司‡　　渡部 広一‡
Ryotaro Mekata　　Seiji Tsuchiya　　Hirokazu Watabe

1　はじめに

近年，情報技術の発展によりユーザは大量の情報を入手可能となった．その一方で，求める情報を的確に選択することが困難となっている．情報を的確に選択する手段の一つとして表形式による要約が挙げられる．表形式で要約することで，文章形式での要約と比べて複数の情報を比較しやすくなり，適切な情報を選択しやすくなると考えられる．そのため，本稿では表形式での要約を扱う．表形式での要約手法は西口らによって，ニュース記事の表形式要約システム[1](以降既存システム)が提案されている．このシステムにおいて，表に出力する項目や単語はそれぞれ 3.3 節や 3.4 節で述べるように獲得されるが，獲得した項目とそこに格納される単語が適切に対応せず，また，共通する項目を持つ単語が少ないため，表には空白が多く生成されていた．そこで本稿では出力する項目の精錬，単語の獲得手法を改善することにより，既存システムの精度向上を目指す．

2　関連技術

2.1　MeCab

MeCab[3]とは，入力された文に対して形態素解析を行うシステムである．形態素解析とは，自然言語で書かれた文の意味を持つ最小の単位である形態素の列に分割し，それらの品詞を判別することである．

2.2　CaboCha

CaboCha[4]とは係り受け解析を行うシステムである．係り受けとは語の間にある修飾-被修飾の関係である．文の係り受けの関係を句・文節を単位として解析する．

3　既存システム

1 つの記事の内容を行，複数記事の内容に共通する項目を列としてまとめて出力する．以下に既存システムが動作する手順を 3.3 節以外は記事 1 を例に用いて示す．

3.1　ニュース記事の分野，見出し，本文の入力

初めに複数記事の入力を行う．入力するのは同分野の記事とし，各記事の見出しと本文，それらの記事の分野を入力する．既存システムではスポーツ，政治，災害などの 7 の分野を設定している．入力記事の例を図 1 に示す．

分野：災害
見出し：JR 東北線で運転見合わせ　西川口駅で人…
本文：京浜東北線は、西川口駅での人身事故の影響…

図 1　入力記事

3.2　ニュース記事から単語，品詞の取得

ニュース記事に対して MeCab による形態素解析を行い，単語とそれぞれの品詞を取得する．形態素解析の結果において，名詞の後に名詞が続く場合は複合語と判断し，1 つの名詞として扱う．単語，品詞を取得する例を図 2 に示す．

見出し：JR 京浜東北線－名詞，で－副詞，運転－…
本文：京浜東北線－名詞，は－助詞，西川口駅－…

図 2　単語，品詞の取得の例

3.3　初期項目と初期名詞，項目パターンの決定

記事から取得した単語より，初期項目に格納する名詞を選択し，それを初期名詞に設定する．初期項目とはあらかじめ分野別に目視で設定された項目，項目パターンとは初期項目に付随する項目群のことを指す．シソーラスを用い，見出し文から取得した名詞の上位ノードを取得する．初期項目の候補が，取得した上位ノードに存在する場合はその名詞を初期名詞とする．初期項目の候補が複数存在する分野の場合，出力する表の初期項目は，入力された記事の中に最も多く存在するものとする．複数記事の初期項目，初期名詞，項目パターンの取得例を図 3 に示す．

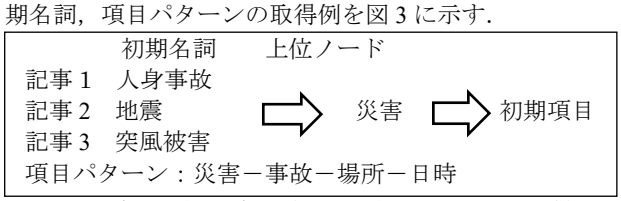

図 3　初期項目，初期名詞，項目パターンの取得

3.4　係り受け解析を用いた重要語の取得

CaboCha を用いて，各記事の本文において前節で決定した初期名詞が修飾している語と修飾されている語を重要語として取得する．ここで，重要語とは表に出力する単語のことを指す．図 4 に係り受け解析による重要語取得の例を示す．

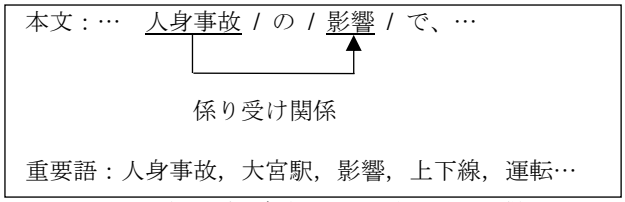

図 4　係り受け解析による重要語の取得

3.5　重要語を格納，表出力

前節で取得した重要語より，出力する表の項目パターンに対応する語を選択し，表を出力する．各重要語の上位ノードをシソーラスより取得し，項目パターン内の項目と一致するノードが存在する場合，その語を項目に対応する内容であると判断し，表に格納する．重要語に対応する項目が項目パターンに存在しない場合，シソーラスにより，その単語の上位ノードに存在する名詞を新たな項目とする．その後，新たな項目は表の右端に生成し，その項目に対応する重要語を格納する．図 5 に既存システムによって生成される表の例を示す．

表 1　既存システムの出力結果

災害	事故	場所	日時	…
人身事故		大宮駅		…
地震		大阪府北部	15 日午前 0 時頃	…
突風被害		高知県内	10 日午前 2 時頃	…

†同志社大学大学院理工学研究科
Graduate School of Science and Engineering, Doshisha University
‡同志社大学理工学部
Faculty of Science and Engineering, Doshisha Universit

3.6 既存手法の問題点

既存システムでは取得した単語ごとにシソーラスを用いて取得しているため，共通した項目を持つ単語がほとんど存在しなかった．また，単語は本文の係り受け解析により取得しているため，要約に必要と思われる単語が適切に取得できていなかった．

4 提案システム

3.6 節で述べた問題点を解決するため，提案システムでは，項目パターンの精錬，および単語の取得手法を新しくすることにより，項目と単語の関係がより適切である表を生成する手法を提案する．以下に 3.2 節で挙げた記事と同じものを使用して，具体的な内容を示す．

4.1 見出し・本文の単語を利用した重要語の取得

記事の見出しに用いられている名詞の単語はその記事において重要なものであると考えられるため，これを利用し，記事の見出しと本文の名詞の単語のうち，部分一致した単語，もしくは完全に一致した単語を記事の重要語として取得する．ここで日時などの日付に関連するものは，見出しで取り扱われることがないため，既存システムと同様にして取得する．取得例を図 5 に示す

見出し：JR 京浜東北線で運転見合わせ　西川口駅…
本文：京浜東北線は、西川口駅での人身事故の影響…
⇩
京浜東北線，西川口駅　…

図 5　単語取得例

4.2 新たな項目パターンの生成

新しく作成する項目パターンは，いつ（時間），どこで（場所），だれが（人物）の要素で構成する．また，全分野共通項目として「メイントピック」という項目を先述した項目パターンとは別に生成する．メイントピックには項目パターンに格納されなかった単語が入り，複数個存在する場合はそれらの単語を重ねて，まとめて「メイントピック」に格納する．また 4.1 節で入力記事が同様の内容でまとめられていると判断された場合，その内容に対応した項目パターンを生成する．

既存システム：災害－事故－場所－日時
提案システム：災害・事故・事件－場所－月・日
－メイントピック

図 6　項目パターン

5 評価

本研究で提案した手法を用いて表形式要約システムの精度評価を被験者 4 人で行った．

5.1 評価方法

朝日新聞社の Web ニュースサイト[5]からニュース記事の分野 7 件，記事 65 件を取得し，テストセットとして扱い．既存システムと提案システムによって表を出力する．出力される表の評価基準を表 1 に示す．

表 1　評価基準

	4 点	3 点	2 点	1 点
内容の理解	○	○	○	×
項目，重要語の量	○	○	×	×
項目名と項目の中身	○	×	×	×

記事の項目を 4 人の合計 16 点満点の内，14 点以上を◎，13 点から 11 点を○，10 点から 8 点を△，7 点以下を×として評価した．

5.2 評価結果

評価結果を表 2 に示す．

表 2　評価結果

	◎（%）	○（%）	△(%)	×(%)
既存システム	0	7.1	18.6	74.3
提案システム	6.2	23.1	1.5	69.2

6 考察

表 1 より既存システムと比べて，提案システムは◎の割合が 6.2%増加，×の割合が 4.9%減少している．3 章で挙げた記事例と同じものを用いて，提案システムによって出力した結果を表 3 に示す．

提案システムで新しく設定した項目パターンにより，表に格納される単語は互いに多くの共通項を持つことができた．しかし，同時に項目メイントピックに多くの単語が格納されてしまい，出力される表は肥大化した．このことから，項目パターンの項目数は不十分であるため，分野単位で必要な項目を探すのではなく，分野内の記事単位で探すことで，より多く単語間に共通する項目を設定することができ，システムの精度向上が期待できる．

また，取得した単語においては，概ね記事の内容が理解できるレベルで取得することができたが，取得した単語すべてを出力するため，項目内に多くの単語が羅列され，表 3 のように出力が文と変わらない結果となる場合が見られた．そのため，取得した単語の中でも重要な情報のみを出力する手法を考案する必要があると思われる．

表 3　提案システムの出力結果

災害・事故・事件	場所	月・日	メイントピック
運転見合わせ			京浜東北線・上下線・高崎線・宇都宮線・西川口駅・人身事故
地震	大阪府北部	15 日午前 0 時 13 分頃	大阪府北部・京都府南部・震度 3・震源・奈良県
突風被害	高知県東部	10 日午前 2 時頃	高知県東部・高知県内・被害・農業用ハウス・倒壊・屋根・損傷

7 おわりに

本稿では，表出力する項目や単語の取得手法を改善するによるニュース記事を表形式に要約する分類手法を提案した．これにより，項目と単語の関係がより適切なものになり，また，共通する項目内に単語が適切に格納される表を出力することが可能となった．その結果，既存システムと比べて◎が 6.2%増加，×が 4.9%減少した．

謝辞

本研究の一部は，JSPS 科研究費 16K00311 の助成を受けて行った．

参考文献

[1] 西口駿祐，芋野美紗子，土屋誠司，渡部広一：“ユーザの要求に応じたニュース記事の表形式要約”，情報科学技術フォーラム FIT2011，E-006，pp.207-208，2011.
[3] Mecab,”MeCab:Yet Another Part-of-Speech and Morphological Analyzer”，http://taku910.github.io/mecab/.
[4] 松本裕治：“形態素解析システム「CaboCha」”，http://chasen.naist.jp/chaki/t/2005-08-29/doc/
[5] “asahi.com：朝日新聞社の速報ニュースサイト”，http://www.asahi.com/